2Ш

УДК 004.652.4+004.827

***Н.Б. Шаховська*[1]*, Д.І. Угрин*[2]**
[1] НУ «Львівська політехніка», м. Львів, Україна
[2] Буковинський університет, м. Чернівці, Україна
natalya233@gmail.com, ugrind@mail.ru

Організація побудови просторів даних туристичної сфери

У статті описано методи отримання, інтеграції та завантаження даних у сховищах даних туристичної сфери. Побудовано модель простору даних туристичної сфери.

Вступ

Останнім часом на розвиток туристичного бізнесу великий вплив мають Інтернет-технології і все частіше у всесвітній павутині можна знайти різноманітні сайти (сторінки), присвячені розвитку туристичної індустрії, туристичним фірмам, агентствам, а також санаторіям, пансіонатам, базам відпочинку та готелям.

Найбільш ефективно використовують переваги Інтернету туристичні фірми та агентства для реклами та надання туристичних послуг. Тому не дивно, що саме продаж туристичних послуг (продаж турів, бронювання авіабілетів, готелів тощо) знаходиться в першій п'ятірці за об'ємами продажу в Інтернеті.

Щодо України, то наразі тут туристичні фірми найбільш активно використовують Інтернет як альтернативний рекламний засіб для своїх послуг. Причинами такої ситуації є розрізненість та неузгодженість даних між туристичними агентствами, органами управління та органами соціальних досліджень, використання різних систем збереження таких даних, відсутність єдиних методів їх інтеграції та опрацювання.

Розроблені на сьогодні підходи до інтеграції даних за своєю функціональністю поділяються на два типи: інтеграції веб-застосувань та інтеграції на основі сховищ даних (з утворенням локального сховища даних). Проте специфіка сфери туризму, а саме:

– наявність великої кількості джерел даних, інформація у яких є різною структурою, не виключає протиріччя та суперечливості інформації;
– наявність великої кількості моделей зберігання джерел даних (реляційні бази даних (РБД), сховища даних (СД), структуровані текстові файли, електронні таблиці, статичні та динамічні веб-сайти тощо);
– відсутність стандартів найменування об'єктів та суб'єктів туристичної галузі;
– ієрархічне впорядкування об'єктів туристичної галузі та агрегування інформації у ході передачі її до верхніх рівнів ієрархії,

вказує на те, що для врахування інформації, отриманої від усіх об'єктів туристичної галузі, необхідно поєднати обидва типи інтеграції.

Можливість об'єднання обох типів інтеграцій дає нова методика зберігання та опрацювання даних – простір даних. Перші роботи з просторів даних з'явилися у 2005 р. (M. Franklin, A. Halevy, D. Maier, D. Kossmann, J.-P. Dittrich). На сьогодні ці роботи мають описовий характер та вказують лише на проблеми, які повинен вирішити простір даних.

Концепція простору даних (ПД) припускає, що учасники простору – зовнішні за відношенням до системи обробки даних, адміністративно-розподілені і семантично гетерогенні джерела даних, – можуть співіснувати з деякою необхідною мірою зв'язності: від простого переліку цих джерел до серйозної БД, що об'єднує їх відповідно до деякої схеми. При цьому концепція ПД передбачає можливість моделювати будь-який вид зв'язку між учасниками.

Об'єктом дослідження є суб'єкти та об'єкти туристичної сфери, зв'язки між ними та інформація, якою вони обмінюються.

1. Актуальність роботи

Простір даних DS – це множина даних, поданих у різних моделях (баз даних **DB**, сховищ даних **DW**, статичних веб-сторінок **Wb**, неструктурованих даних **Nd**, графічних та мультимедійних даних **Gr**), локальних сховищ та індексів (**ODW**), а також засобів інтеграції (**Int**), пошуку (**Se**) та опрацювання інформації (**Wo**), об'єднаних середовищем управління моделями (**EM**) [2]:

$$\mathbf{DS} = < \mathbf{DB}, \mathbf{DW}, \mathbf{ODW}, \mathbf{Wb}, \mathbf{Nd}, \mathbf{Gr}, \mathbf{Int}, \mathbf{Se}, \mathbf{Wo}, \mathbf{EM} >.$$

У сучасному світі, що прагне до глобалізації відносин, період між надходженням даних і ухваленням рішення неухильно скорочується. Усе більше процесів вимагає аналізу в режимі реального часу, коли будь-яка затримка є критичною. І в таких ситуаціях без допомоги ефективного інструментарію обробки даних, побудованого на інформаційних технологіях, туристичному операторові не обійтися.

Простір даних дозволяє туристичному операторові самостійно розподіляти свої тимчасові ресурси. З використанням процедур з'являється можливість багаторазово використовувати власну працю, яку затрачено на пошук рішення в аналогічній ситуації, узагальнювати й аналізувати власний досвід, обмінюватися ним з іншими туристичними організаціями. У підсумку повинна з'являтися можливість заощаджувати час туристичних операцій, які затрачено на прийняття складних рішень з більшим числом невідомих. А для цього нам потрібні механізми структурування, агрегування, ефективного пошуку й аналізу даних, які й повинен надавати простір даних.

2. Постановка задачі

Опишемо об'єкти туристичної сфери, інформація з яких повинна опрацьовуватися іншими об'єктами туристичної сфери:

Місцеві органи управління – надають інформацію про відпочинкові, рекреаційні та оздоровчі ресурси, а також правила їх експлуатації; шляхи сполучення, особливості місцевості тощо.

Туристичні агентства – надають інформацію про себе, про свої послуги.

Адміністративні одиниці – описуються через інформацію місцевих органів управління, а також через відгуки попередніх відвідувачів.

Особа (відпочивальник) – надає інформацію про себе, про умови, які він хоче отримати, ціни тощо.

Залежно від типу об'єкта інформація може зберігатися у різних моделях та надходити з різних джерел.

Туристичне агентство – база даних, динамічний веб-сайт з базою даних, розміщеною на веб-сервері.

Адміністративна одиниця – сховище даних.

Особа – веб-сайт, база даних, текстові дані тощо.

Відпочинковий ресурс – база даних, веб-сайт.

Об’єкти туристичної галузі та задачі, що перед ними ставляться, подані на рис. 1.

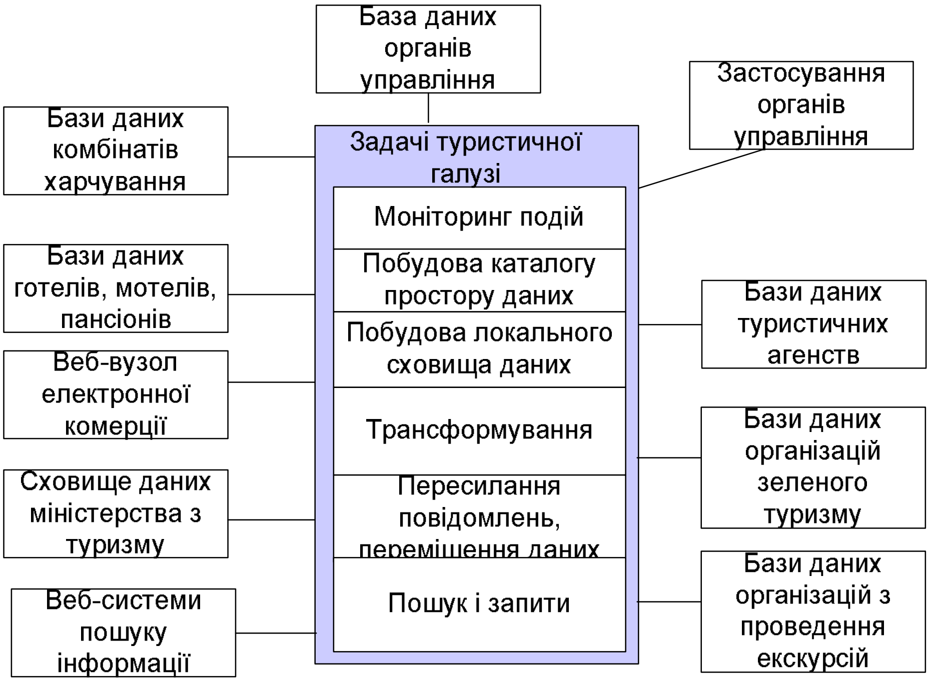


Рисунок 1 – Об’єкти туристичної сфери та задачі, що перед ними ставляться

Розглянемо сфери застосування інформаційних технологій у туристичній галузі. Вони подані на рис. 2.

Більшість користувачів туристичної сфери у своїй роботі використовують досить обмежені набори даних, які варіюються залежно від типу діяльності. У принципі для ефективної організації власного простору туристична організація повинна мати можливості, подані на рис. 3.



Рисунок 2 – Сфери застосування інформаційних технологій в туристичному бізнесі



Рисунок 3 – Схема організації власного простору для туристичної організації

Розглянемо ці можливості більш докладно.

1. Вибір джерел даних.

Дані можуть вибиратися з будь-яких доступних джерел. Це може бути Інтернет, файлові джерела (локальні й мережні), структуровані бази даних і т.д. При цьому можна брати не все джерело, а тільки його частину, попередньо обмежуючи масив доступних для відбору даних (наприклад, не весь локальний диск, а тільки папку, не весь Інтернет, а тільки певний сайт, не всю базу даних, а тільки певну таблицю).

2. Відбір даних із джерел.

Для відбору даних у першу чергу потрібен потужний і функціональний механізм пошуку. Користувач повинен мати можливість відбирати все, що завгодно, як за ключовими словами, так і за допомогою різних засобів інтелектуального пошуку.

3. Збереження результатів відбору.

Відібрані дані необхідно зберегти, причому вже на цьому етапі користувач повинен мати можливість провести попередню структуризацію отриманих даних. Для цього користувачеві надається можливість створювати структуровані набори – бази даних, за допомогою яких, відповідно до своїх запитів, він може задавати структуру даних, що зберігаються. За бажанням користувача бази даних можуть бути певним чином проіндексовані – для спрощення подальшої роботи з ними.

4. Робота з даними.

Збережені дані потрібно опрацювати. Для цього необхідно мати повний набір можливостей для роботи з даними, включаючи пошук, перегляд, редагування й перетворення. Працювати можна як з окремими елементами бази даних, так і з самими базами даних, застосовуючи до них різні операції перетворення, пов’язані з обчисленнями, об’єднаннями, порівняннями й т.п. Перетворені бази даних також зберігаються.

5. Збереження сценаріїв.

Вся зроблена робота зі знаходження рішення повинна зберігатися у вигляді процедури, за допомогою якої в майбутньому можна проаналізувати виконану роботу. Процедури дозволять згодом прийняти рішення під час пошуку можливих альтернатив, узагальнити результати, використати для рішення аналогічних завдань.

6. Обмін даними.

Для обміну даними повинні використовуватися різні механізми, наприклад, архівування та передача мережами, вивантаження в зовнішній формат, публікація тощо.

3. Основний матеріал

У просторі даних повинна використовуватися вся інформація, що потрібна для конкретної туристичної організації, незважаючи на формат і місце розташування цієї інформації, а також повинен моделюватися розвинутий набір зв’язків між репозитаріями даних. Тому на логічному рівні простір даних представляється як набір *учасників* і *зв’язків*. Разом з неоднорідністю вмісту простору даних виникає потреба в підтримці декількох стилів доступу до даних. ПППД допускає багато різних режимів взаємодії, і потрібна гранична спільність, щоб допустити застосування різних служб до різних типів вмісту.

Архітектуру простору даних розділимо за рівнями (рис. 4).

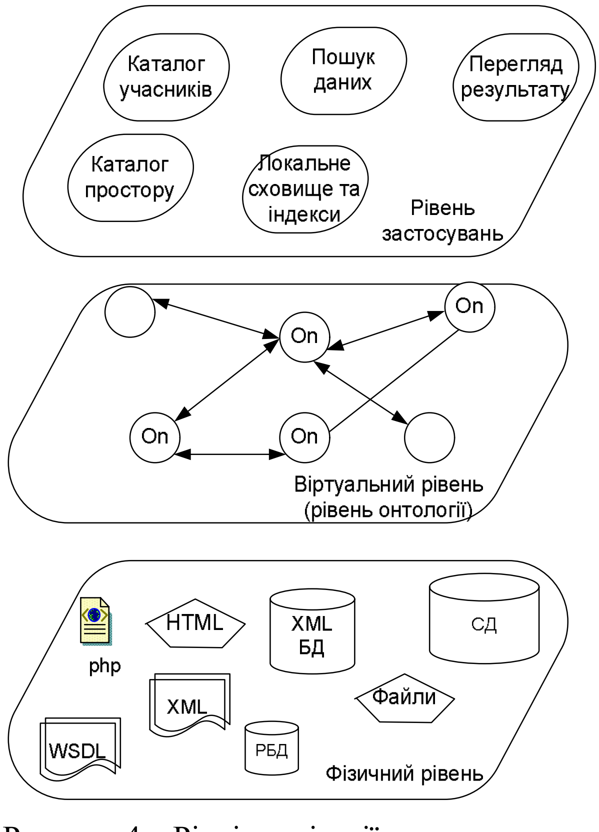


Рисунок 4 – Рівні реалізації простору даних

Базовою службою простору даних є каталогізація елементів даних, отриманих від учасників. **Каталог** – це реєстр ресурсів даних, що утримує найбільш загальну інформацію про кожний з них: джерело, ім’я, місце розташування в джерелі, розмір, дата створення й власник і т.д. Каталог є інфраструктурою для більшості інших

сервісів просторів даних, але він також може підтримувати базовий користувацький інтерфейс перегляду простору даних:

Metadata(DB, DW, Wb, Nd, Gr) ⇒ Cg.

Він не тільки містить описову інформацію (тобто виконує роль метаданих), але й зберігає для кожного учасника схему джерела, статистичні дані, швидкість зміни, точність, можливості відповідей на запити, інформацію про власника і дані про політику доступу і підтримку конфіденційності. Оскільки джерела простору даних фізично не переносять у нього інформацію та можуть обмінюватись між собою інформацією, то у каталозі необхідно зберігати дані і про зв'язки між джерелами.

Зв'язки у каталозі можуть зберігатися у вигляді:
– метаданих;
– перетворень запитів;
– графів залежності;
– текстових описів тощо.

Призначення каталогу можна охарактеризувати таким чином:
– визначення семантично близьких назв об'єктів. Прикладом може бути WordNet. Це велика семантична лексична база даних англомовних термінів, синонімів, прийнятих скорочень, зв'язків між синонімами;
– визначення користувачів та їх прав доступу (привілеїв);
– визначення джерел даних, структур даних та зв'язків між ними.

Двома основними службами, які будуть підтримуватися в ПППД, є пошук і запити даних. Пошук є основним механізмом роботи кінцевих користувачів з більшими колекціями незнайомих даних. Пошук менш вимогливий, ніж запити даних, оскільки він заснований на подібності, наданні кінцевим користувачам ранжованих результатів і підтримці інтерактивного вдосконалювання. ПППД повинні дозволяти користувачам задавати пошуковий запит й ітераційно його вдосконалювати, якщо це доречно, до вигляду запиту в стилі бази даних. Ключовий принцип просторів даних полягає в тому, що пошук повинен бути застосовуваним до всього вмісту простору даних, незалежно від форматів даних. Універсальні можливості пошуку й запитів повинні поширюватися на метадані. У користувачів повинні бути можливості знаходження необхідних джерел даних і одержання інформації про їхню складність, коректність і актуальність.

Відповідно до перерахованих служб, виділяються наступні архітектурні компоненти ПППД.

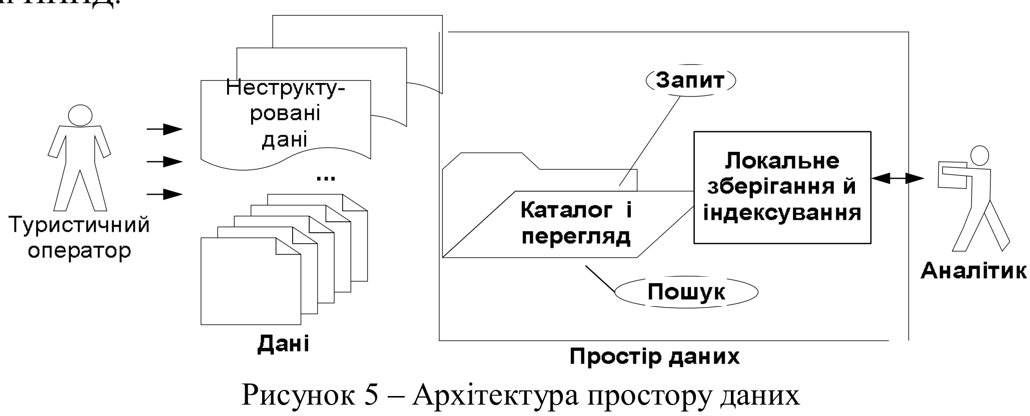


Рисунок 5 – Архітектура простору даних

Каталог і перегляд. Каталог містить інформацію про всіх учасників простору даних і про зв'язки між ними. Для кожного учасника каталог повинен включати схему джерела, статистичні дані, швидкість зміни, точність, можливості відповідей на

запити, інформацію про власника й дані про політику доступу й підтримку конфіденційності. Зв'язки можуть зберігатися у вигляді перетворень запитів, графів залежності, а іноді навіть у вигляді текстових описів.

По можливості в каталозі повинен міститися базовий реєстр елементів даних по кожному учаснику: ідентифікатор, тип, дата створення й т.д. Тоді в ньому можна підтримувати базову можливість перегляду об'єднаного реєстру всіх учасників. Інтерфейс перегляду можна використати для відповідей на питання користувачів про наявність або відсутність елемента даних або визначення того, які учасники зберігають документи даного типу.

Навколо каталогу ПППД повинне підтримуватися середовище керування моделями, що дозволяє створювати нові зв'язки й маніпулювати існуючими зв'язками (наприклад, поєднувати або інвертувати відображення, зливати схеми й створювати єдині подання декількох джерел).

Пошук і запити. У користувачів повинна бути можливість запиту будь-якого елементу даних, незалежно від його формату й моделі даних.

Щодо носіїв простору даних виконуються такі операції з множини **Se**:

1. *Запит про довільні дані* $\mathrm{Se}_{\mathrm{simple}}$ – у користувачів повинна бути можливість запиту будь-якого елементу даних, незалежно від його формату і моделі даних. Здійснюється на основі ключових слів key_word та каталогу даних **CG**. Спочатку ПППД повинні підтримувати для кожного учасника запити за ключовими словами. У міру того, як ми одержимо більше інформації про учасника, ми повинні поступово почати підтримувати складніші запити. Система повинна підтримувати плавне перемикання між запитами за ключовими словами, переглядом і структурованими запитами. Зокрема, при видачі відповідей на запит за ключовими словами (або на структурований запит) повинні пропонуватися додаткові інтерфейси запитів, що дозволяють користувачу удосконалити свій запит:

$$\mathrm{Se}_{\mathrm{simple}} : \sigma_{key_word(Cg)}.$$

2. *Стуктуровані запити* будуються з використанням SQL та подібних мов. За допомогою каталогу визначається, чи джерело, у якому здійснюватиметься пошук, містить структуровану інформацію. Якщо це так, то виконується запит безпосередньо до джерела даних. В іншому випадку запит продовжує виконуватись за каталогом даних у вигляді пошуку ключових слів. Запити в стилі баз даних повинні підтримуватися на основі загальних інтерфейсів (тобто схем-посередників), що забезпечують доступ до декількох джерел, або вони можуть адресуватися конкретному джерелу даних (з використанням його власної схеми) з наміром отримання відповідей і від інших джерел (як в системах керування одноранґовими даними – Peer-Data Management System) [3]. Запити можуть формулюватися на різноманітних мовах (і на основі різних моделей даних), і вони повинні, якщо можливо, найкращим чином переформулюватися на інші моделі даних і схеми, забезпечуючи точні і приблизні семантичні відображення:

$$\mathrm{Se}_{\mathrm{structured}} : \sigma_{key_word(Cg)}, \sigma(Source).$$

3. З*апити до метаданих* повинні забезпечувати можливості:

– отримання даних про джерело відповіді та місцезнаходження джерела;
– визначення елементів даних в просторі даних, що можуть залежати від заданого елементу даних, і підтримка гіпотетичних запитів;
– визначення рівня невірогідності відповіді.

Схема пошуку інформації за запитом подана на рис. 6.

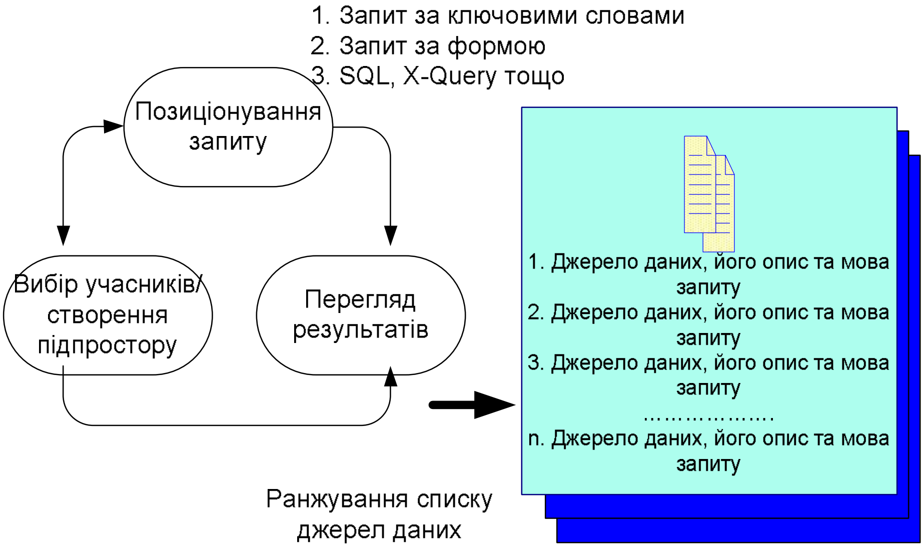


Рисунок 6 – Схема пошуку інформації у просторі даних

Локальне зберігання й індексування. У ПППД повинен бути компонент зберігання й індексування, що забезпечує наступні можливості: створення запитуваних асоціацій між об'єктами даних, отриманих від різних учасників; удосконалювання доступу до джерел з обмеженими власними засобами доступу; можливість виконання деяких запитів без доступу до реального джерела даних; підтримку високого рівня доступності й відновлення.

Засоби індексування повинні мати високий рівень адаптивності до неоднорідних середовищ. Для вхідних даних необхідно, щоб приймалося будь-яке елементарне значення, що зустрічається в просторі даних, і повинні видаватися координати всіх об'єктів даних, у яких є таке значення, і ролі кожного його входження. Важливими аспектами індексу є те, що, по-перше, він визначає інформацію *для всіх* учасників, коли деякі значення входять у кілька джерел даних. По-друге, індекс повинен справлятися з різнорідністю посилань на об'єкти бази.

Компонент розкриття. Призначення цього компонента полягає у виявленні учасників у просторі даних, створенні зв'язків між ними й наданні допомоги адміністраторам при вдосконалюванні й посиленні цих зв'язків. Виявлення учасників може відбуватися в декількох формах, наприклад, у формі обходу довідкової структури, починаючи від кореня, або у формі пошуку координат всіх баз даних у корпоративній мережі. Компонент повинен виконувати початкову класифікацію учасників на основі їхніх типів і вмісту.

Після розкриття учасників система повинна забезпечити середовище для напівавтоматичного створення зв'язків між учасниками й удосконалювання та підтримки існуючих зв'язків. Цей процес включає знаходження пар учасників, які, ймовірно, повинні бути зв'язані один з одним, і пропозицій зв'язків, які потім перевіряються та уточнюються людиною. Компонент розкриття повинен здійснювати моніторинг умісту простору даних, щоб можна було згодом запропонувати нові зв'язки.

Компонент розширення джерел. У деяких учасників можуть бути відсутні істотні функції керування даними. У ПППД повинні бути засоби наповнення такого учасника додатковими можливостями, такими, як схема, каталог, пошук за ключови-

ми словами й моніторинг відновлень. Може виявитися необхідність забезпечувати ці розширення за відділеннями чи напрямками, оскільки можуть бути реальні додатки або потоки даних, розраховані на наявні формати або довідкові структури.

Висновки

У статті описано простір даних туристичної сфери, що забезпечує взаємодію між джерелами інформації, поданої за допомогою різних моделей даних, з різними методами подання та опрацювання. Наукова новизна полягає в уведенні формального опису простору даних та окреслення його основних задач. Практична цінність полягає у побудові простору даних туристичної сфери, виділенні основних об'єктів та учасників. Подальші дослідження стосуватимуться формалізації методів інтеграції даних та пошуку неструктурованих, напівструктурованих та строго структурованих даних.

Література

1. Franklin M. From Databases to Dataspaces: A New Abstraction for Information Management [Електронний ресурс] / M. Franklin, A. Halevy, D. Maier // ACM SIGMOD Record 34. – 2006. – №. 4. – Режим доступу : http://www.sigmod.org/sigmod/record/issues/0512/p27-article-franklin.pdf.
2. Шаховська Н.Б. Простори даних: поняття та призначення / Н.Б. Шаховська // Матеріали конференції CSIT-2007. – Львів, 2007. – С. 269-277.
3. Halevy Alon. Why Your Data Won't Mix [Електронний ресурс] / Alon Halevy // ACM Queue, 3. – 2005. – № 8. – Режим доступу : http://www.acmqueue.com/modules. php?name= Content&pa=showpage& pid= 336

Н.Б. Шаховская, Д.И. Угрин
Организация построения пространств данных туристической сферы
В статье описаны методы получения, интеграции и загрузки данных в хранилищах данных туристической сферы. Построена модель пространства данных туристической сферы.

N.B. Shahovska, D.I. Ugryn
An order, methods and facilities of getting, concordance, integration of information, creation of operative depositories of information and load of information, is in-process worked out in a central depository. The dataspase model is build. The maim problems of tourism sphere are described.

Стаття надійшла до редакції 16.03.2009.